# Generation of Eye Contact Image Using Depth Camera for Realistic Telepresence

Sang-Beom Lee and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST),
Gwangju, 500-712, Republic of Korea
E-mail: {sblee, hoyo}@gist.ac.kr

*Abstract*— **In this paper, we present an eye contact system for realistic telepresence using a depth camera that utilizes an infrared structured light. In order to generate the eye contact image, we capture a pair of color and depth video and separate the foreground single user from the background. Since the raw depth data includes several types of noises, we perform a joint bilateral filtering method to reduce the noise. Then, we apply a discontinuity-adaptive depth filter to the depth map to reduce the disocclusion region. From the color image and the preprocessed depth map, we construct a three-dimensional model of the user at the virtual viewpoint. The entire system is implemented through GPU-based parallel programming for real-time processing. Finally, we obtain the gaze-corrected user. Experimental results show that the proposed system is efficient in realizing eye contact and provides more realistic experience.**

## I. INTRODUCTION

3D video, which contains a color video and the corresponding depth video, can be obtained by means of several sensing techniques. They are mainly classified into two categories: passive sensor-based method and active sensor-based method. The former calculates depth data indirectly by inferring the correlation of 2D images captured by two or more cameras. Stereo matching is a well-known process based on this method [1]. The latter, on the other hand, makes use of various types of sensors, such as laser, infrared ray (IR), or light pattern, to directly obtain depth information from the real 3D scene. Mostly, depth cameras, structured light sensors, and 3D scanners are employed in this method [2]. Conventionally, the active sensor-based method has not been a wise choice to reach consumers, due to the huge cost of depth cameras. However, recently, manufacturers have introduced cheaper and smaller depth cameras, allowing wide use in 3D home game and multimedia [3]. Hence, the active sensor-based method is spotlighted as one of the powerful technologies in 3D content production.

Using the color image and its corresponding depth map, the virtual image can be synthesized by using depth image-based rendering (DIBR) [4]. Generally, in real 3D scenes, the depth map can be regarded as 3D information. The procedure of virtual image generation is as follows. First, using camera geometry and depth map, all pixels of the color image in the original viewpoint are back-projected to the world coordinate. Subsequently, the points in the world coordinate are re-projected to the image plane at the virtual viewpoint. Such a procedure is called "3D warping" in computer graphics literature [5].

DIBR is widely used in multimedia industry. Especially, telepresence is a set of technologies which allows a person to communicate with others while giving the impression of an in-person conversation. The eye contact issue, one of the most challenging issues in telepresence, is becoming a hot topic for researchers. Although many algorithms have been proposed, the issue still remains problematic. However, recent studies have effectively used DIBR, making telepresence more realistic. The improved telepresence has reached consumer solutions, e.g., videoconferencing, education, and virtual reality. The recent proposals set multiple cameras mounted near the display while depth data are estimated from multi-view images [6][7]. Afterward, virtual view synthesis is performed by warping captured images. However, these systems require complex setup and complicated algorithms. Especially, since the performance of stereo matching is very sensitive to scene environment, we cannot expect feasible view synthesis results.

In this paper, we propose an eye contact system using only one depth camera for realistic telepresence. The main objective of our work is to create a practical and stable eye contact system with the aid of a depth camera that utilizes infrared structured light pattern. The system is simply configured by a depth camera mounted on the top of the display. The main processes of the proposed system, which require massive computational time, are implemented via GPU-based parallel programming for real-time processing. In addition, to improve the depth quality, the proposed system includes three depth preprocessing algorithms: foreground/ background separation, joint bilateral filtering, and discontinuity-adaptive filtering.

## II. SYSTEM FRAMEWORK

Fig. 1 shows the proposed system framework to generate the virtual eye contact image. Initially, the depth camera captures the color image and its corresponding depth map. Then, a single user is separated from the background in the preprocessing step. The depth camera cannot perfectly obtain the depth values for the entire scene due to sensor noises and occlusion regions. Therefore, we conduct joint bilateral filtering to estimate the unoccupied regions. In order to reduce disoccluded region, newly exposed areas after 3D warping,

we apply the discontinuity-adaptive filtering operation. With the preprocessed depth map and the color image, we project the entire pixel of the color image to the world coordinate. Then, we construct a 3D human model that is composed of triangular meshes using the points of the world coordinate. After synthesizing virtual image and filling remaining holes, finally we can obtain the eye contact image.

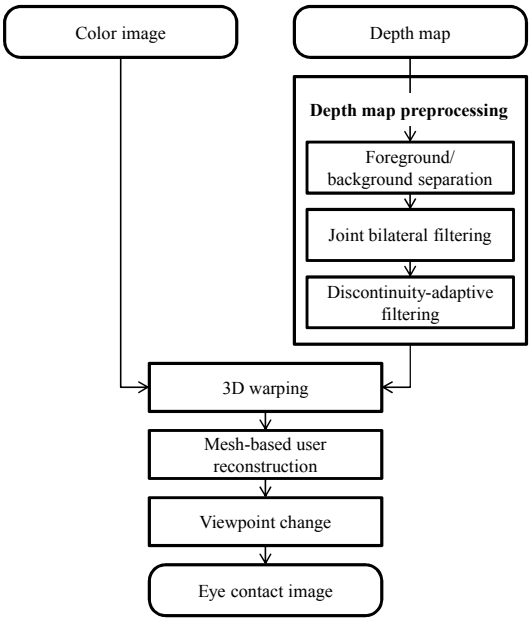



Fig. 1 Specification of the proposed system.

## III. DEPTH MAP PREPROCESSING

### A. Foreground/Background Separation

In order to identify the single user in front of the depth camera, we separate the user from the background. One of the powerful characteristics of the depth camera is the ability to change parameters of depth range. This parameter can be set to include or remove certain objects, depending on the specific need of users.

(a) Color image and depth map (b) Separated foreground user

Fig. 2 Results of foreground/background separation.

In the proposed scenario, we separate only the foremost user for videoconferencing. The system initially finds the minimum depth value of the scene. Then, we determine the depth range as 1,000 mm from the minimum depth. Fig. 2 shows the results of foreground separation. Fig. 2(a) represents the raw color image and the associated depth map. As shown in Fig. 2(b), we successfully separate the user by using the minimum depth value and depth range. Notice that the depth map in this paper is illustrated by the normalization of depth values from 0 (nearest) to 255 (farthest).

### B. Joint Bilateral Filtering

The depth camera is difficult to find depth values due to its inherent problems such as sensor noise, lost depth information in shiny or black surfaces and occlusion regions. The black empty areas, as shown in Fig. 2(a) and Fig. 2(b), are caused by depth camera errors.

To fill the empty areas, we exploit the iterative joint bilateral filter (JBF) [8]. The key idea of JBF in the proposed system is the use of two Gaussian filters: a range filter of color image and a spatial filter. JBF is defined by

$$D(x,y)=\frac{\sum_{u\in u_p}\sum_{v\in v_p}W(u,v)\cdot D(u,v)}{\sum_{u\in u_p}\sum_{v\in v_p}W(u,v)} \tag{1}$$

$$W(u,v)=\begin{cases}0 & if\ D(u,v)=0\\ g_I(u,v)\cdot f(u,v) & otherwise\end{cases} \tag{2}$$

$$g_I(u,v)=\exp\left\{-\frac{|I(x,y)-I(u,v)|^2}{2\sigma_R^{\,2}}\right\}$$
$$f(u,v)=\exp\left\{-\frac{(x-u)^2+(y-v)^2}{2r^2}\right\} \tag{3}$$

where $u_p=\{x\text{-}r,\ldots,x+r\}$ and $v_p=\{y\text{-}r,\ldots,y+r\}$, while $r$ represents the filter radius. The filter variance and radius are designated as $\sigma_R$=255 and $r$=3 in the proposed system. JBF is only applied to the empty areas and the iteration process is continued still those areas are removed. Fig. 3 shows the results of joint bilateral filtering applied to the depth map.

(a) Color image (b) Filtered depth map

Fig. 3 Results of joint bilateral filtering.

### C. Discontinuity-adaptive Filtering

Fig. 4 shows the 3D warping result for the virtual viewpoint. The empty spaces near the user's neck, shown in

Fig. 4, are newly exposed area. The region is called "disocclusion" in computer graphics literature. Since no texture information exists, we should fill the disoccluded areas for natural view synthesis.

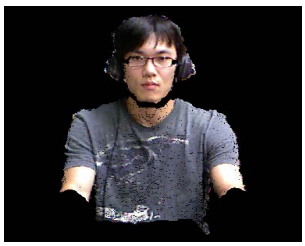


Fig. 4 3D warping result.

The best-known algorithm for disocclusion filling is depth map smoothing by means of a Gaussian low-pass filter. Weakening the sharpness of depth discontinuity is the main advantage, and disoccluded areas are mostly filled from neighboring pixels. As we apply this method to the depth map and perform bilinear interpolation by averaging textures from neighborhood pixels, the disocclusion can be reduced. However, since the depth map is seriously deformed, smoothing out the entire depth values is undesirable.

Therefore, we apply the discontinuity-adaptive filter to the depth map prior to 3D warping [9]. If we analyze the amount of depth discontinuity at object edges and determine the filtering range, we not only minimize depth deformation but also improve the visual quality of the virtual image. The filtered depth at $(x,y)$ in the depth map $D$ is defined by

$$D_{filtered}(x,y) = \alpha(x,y)\cdot D(x,y) + \{1-\alpha(x,y)\}\cdot D_{Gaussian}(x,y) \quad (4)$$

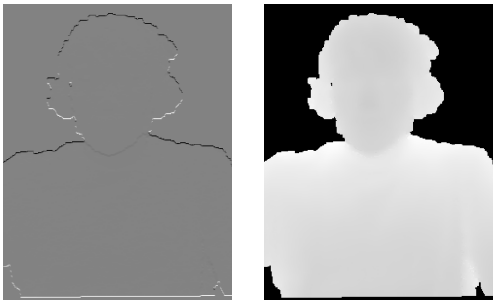
$$\alpha(x+u,y+v) = \begin{cases} \dfrac{|x-u|+|y-v|}{\delta(x,y)} & \text{if } |x-u|+|y-v| < \delta(x,y) \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

$$D_{Gaussian}(x,y) = \sum_{v}\sum_{u} D_{original}(x-u,y-v)\cdot g(u,v) \quad (6)$$

$$g(u,v) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|u-v|^2}{2\sigma^2}\right) \quad (7)$$

where $D_{Gaussian}(x,y)$ represents the filtered depth map by Gaussian smoothing filter. $\delta(x,y)$ denotes the depth discontinuity at $(x,y)$. The range of $u$ and $v$ are $-D(x,y)\le u\le D(x,y)$ and $-D(x,y)\le v\le D(x,y)$, respectively. Moreover, the window size is set to be three times larger than the variance of the filter.

Fig. 5 shows the filtering results of the proposed system. As shown in Fig. 5(a), we extract the edge of the depth map and analyze the depth discontinuity strength. Then, the filtering range is determined by the depth discontinuity near the object boundary. We obtain the smoothed depth map as shown in Fig. 5(b). After we perform 3D warping, remaining holes can be efficiently removed by interpolation using the neighboring pixels. Finally, we obtain the preprocessed depth map.



(a) Discontinuity map (b) Preprocessed depth map

Fig. 5 Results of discontinuity-adaptive filtering.

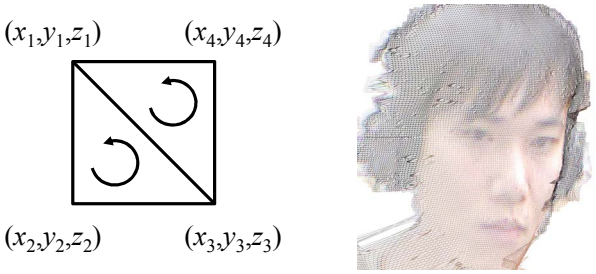
## IV. Virtual View Synthesis

### A. Mesh-based User Reconstruction

In the proposed system, we focus on triangular mesh-based representation of the user with preservation of regularities, e.g., image-based representation. With the point cloud, we reconstruct the 3D user model in the world coordinate using triangular mesh surfaces. Originally, we find four corner pixels in a counter-clockwise direction. Since the four pixels are regarded as four vertices of triangles, which have $x$, $y$, and $z$ coordinates in the world coordinate, we can define these vertices, $v_1$, $v_2$, $v_3$, and $v_4$ as follows:

$$v_1=\{x_1, y_1, z_1\},\ v_2=\{x_2, y_2, z_2\},\ v_3=\{x_3, y_3, z_3\},\ v_4=\{x_4, y_4, z_4\} \quad (8)$$

By using the vertices, we construct two triangles: One is composed of $v_1$, $v_2$, and $v_3$. The other is composed of $v_1$, $v_3$, and $v_4$. Since each vertex possesses its own color, the triangles are filled with interpolated values of three colors. Fig. 6(a) shows the construction of triangles with four neighbor pixels. The triangulation result of the point cloud is represented in Fig. 6(b).




(a) Triangulation with four pixels (b) Mesh triangulation

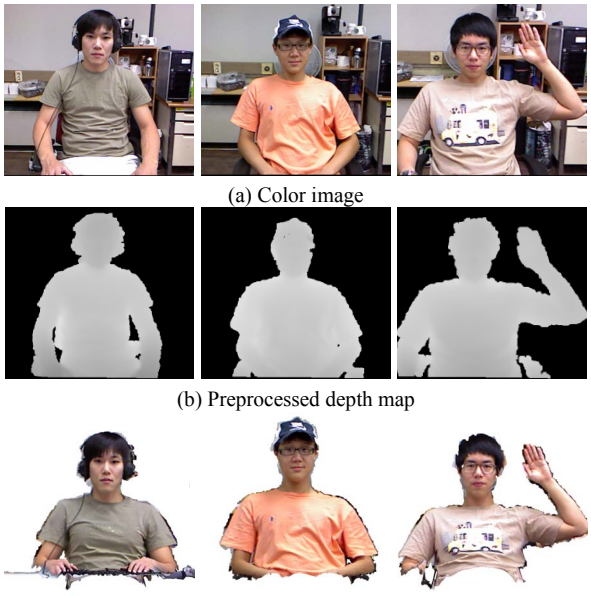
Fig. 6 Result of mesh triangulation.

### B. Viewpoint Change of the User Model

In order to synthesize the eye contact image, we modify the camera viewpoint. By matching the optical axis of the depth camera and user gaze, we display the gaze-corrected user model to the distant user. Finally, the gaze-corrected user is generated, making realistic telepresence is achievable.

## V. EXPERIMENTAL RESULTS

To fabricate the gaze-corrected image, we have constructed the depth camera system, utilizing infrared structured light pattern. The depth camera employs a resolution of 640×480 at 30 frame/s, with depth range dynamically changing. However, due to the long user-camera distance, the depth sensor accuracy is degraded. Therefore, we determined the maximum depth range as 1500 mm. Initially, the optical axis of the camera is toward the user head.

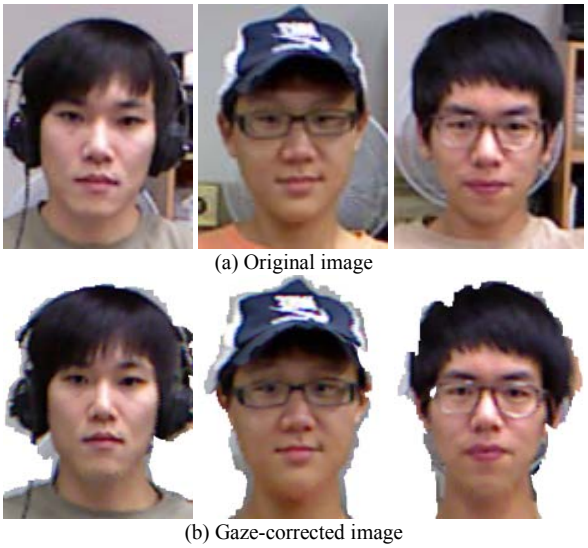


(a) Color image

(b) Preprocessed depth map

(c) Gaze-corrected user model

Fig. 7 Results of gaze-correction.



(a) Original image

(b) Gaze-corrected image

Fig. 8 Zoom-in images of users' face.

Fig. 7 shows the results of gaze-correction. Fig. 7(a) and Fig. 7(b) represent original color images and their preprocessed depth maps, respectively. In Fig. 7(c), despite some errors detected at user heads, we have produced smooth gaze in the images. Moreover, exploiting depth values, we easily separate the users from the complex background. Fig. 8 shows the zoomed-in user faces. Fig. 8(b) displays more precise eye contact compared to Fig. 8(a).

## VI. CONCLUSIONS

In this paper, we have presented a new approach to generate gaze-corrected images using the depth camera system. The proposed system has used several depth map preprocessing techniques: foreground/background separation, joint bilateral filtering, and discontinuity-adaptive filtering. With the color image and its corresponding depth map, we constructed the 3D user model represented via triangular mesh-based representation. Finally, we synthesized the gaze-corrected image by means of viewpoint change. Experimental results have verified that we efficiently realized eye contact. Our system requires only a depth camera and a display, allowing simplicity and flexibility. Thus, our system will be able to fit into various commercial solutions.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0009228).

## REFERENCES

[1] D. Sharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," *IEEE Workshop on Stereo and Multi-Baseline Vision*, pp. 131–140, Dec. 2001.

[2] E. Lee and Y. Ho, "Generation of Multi-view Video Using a Fusion Camera System for 3D Displays," *IEEE Transactions on Consumer Electronics,* vol. 56, no. 4, pp. 2797–2805, Nov. 2010.

[3] L. Xia, C. Chen, and J. K. Aggarwal, "Human Detection Using Depth Information by Kinect," *Computer Vision and Pattern Recognition Workshops*, pp. 15–22, June 2011.

[4] C. Fehn, "Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3-D TV," *SPIE Conference Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93–104, Jan. 2004.

[5] W. R. Mark, L. McMillan, and G. Bishop, "Post-Rendering 3D Warping," *Symposium on Interactive 3D Graphics*, pp. 7–16, April 1997.

[6] O. Schreer, N. Atzapadin, and I. Feldmann, "Multi-baseline Disparity Fusion for Immersive Videoconferencing," *International Conference on Immersive Telecommunications*, pp. 27–29, May 2009.

[7] S. Lee, I. Shin, and Y. Ho, "Gaze-corrected View Generation Using Stereo Camera System for Immersive Videoconferencing," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1033–1040, Aug. 2011.

[8] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint Bilateral Upsampling," *SIGGRAPH'07*, pp. 96–100, Aug. 2007.

[9] S. Lee and Y. Ho, "Discontinuity–adaptive Depth Map Filtering for 3D View Generation," *International Conference on Immersive Telecommunications*, pp. T8(1–6), May 2009.